MS COCO_Datataset_2017: https://cocodataset.org/#download

I downloaded from here: https://www.kaggle.com/datasets/awsaf49/coco-2017-dataset

Directory: Coco 2017
|---annotations
  |---captions_train2017.json (Captions for training images)
  |---captions_val2017.json (Captions for validation images)
|---train2017 (Contains 118K training set images)
|---test2017 (Contains 40.5K training set images)
|---val2017 (Contains 5K training set images)

The overall approach remains the same for both datasets (Flickr8K and COCO), with only minor differences in caption loading—Flickr8K uses plain text files, while COCO uses JSON format. Corresponding adjustments are made for preprocessing.

In Flickr8K, captions are directly available in `Flickr8k.lemma.token.txt` for all 8K images. In contrast, for COCO, we extract captions from the annotation JSON file and generate a `descriptions.txt`, while image features are saved in `features.p` separately for training, validation, and test sets. The tokenizer is built on the training set and reused for validation and testing.

Due to the significantly larger size of the COCO dataset, initial training was performed on a subset of 14K images (more than Flickr8K) for 10 epochs. Subsequently, the model was trained on the full COCO dataset of 118K images for 6 epochs.

Results:

Initial training was performed on 14K images for 10 epochs (with 6.68 million trainable parameters). The model's training loss decreased steadily from **4.69 to 3.46**

17083/17083 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 277s 16ms/step - loss: 4.6908
17083/17083 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 268s 16ms/step - loss: 3.8052
17083/17083 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 270s 16ms/step - loss: 3.6090
17083/17083 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 268s 16ms/step - loss: 3.5236
17083/17083 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 268s 16ms/step - loss: 3.4847
17083/17083 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 267s 16ms/step - loss: 3.4735
17083/17083 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 265s 16ms/step - loss: 3.4682
17083/17083 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 265s 16ms/step - loss: 3.4636
17083/17083 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 265s 16ms/step - loss: 3.4675
17083/17083 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 264s 15ms/step - loss: 3.4687

Validation BLEU scores for the final model (model9.h5) were:
BLEU-1: 0.3241, **BLEU-2: 0.1596**, **BLEU-3: 0.1108**, and **BLEU-4: 0.0486**.

Subsequently, full-scale training was done on all **118K images**, with a vocabulary size of **27,089** and a maximum caption length of **49**, for 6 epochs. Training loss improved slightly from **4.06 to 3.88** (with a small spike at epoch 5). However, BLEU scores on the test set remained modest
144316/144316 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2742s 19ms/step - loss: 4.0683
144316/144316 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2705s 19ms/step - loss: 3.9154
144316/144316 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2691s 19ms/step - loss: 3.8903
144316/144316 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2693s 19ms/step - loss: 3.8892
144316/144316 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2695s 19ms/step - loss: 3.8814
144316/144316 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2814s 19ms/step - loss: 3.9660

Validation Scores:
model_0.h5 BLEU-1: 0.285099 BLEU-2: 0.130661 BLEU-3: 0.071381 BLEU-4: 0.022918
model_1.h5 BLEU-1: 0.241617 BLEU-2: 0.095296 BLEU-3: 0.054147 BLEU-4: 0.018392
model_2.h5 BLEU-1: 0.241866 BLEU-2: 0.098066 BLEU-3: 0.040607 BLEU-4: 0.006897
model_3.h5 BLEU-1: 0.259875 BLEU-2: 0.105982 BLEU-3: 0.054820 BLEU-4: 0.015524
model_4.h5 BLEU-1: 0.264036 BLEU-2: 0.107859 BLEU-3: 0.053310 BLEU-4: 0.015857
model_5.h5 BLEU-1: 0.252894 BLEU-2: 0.104606 BLEU-3: 0.052580 BLEU-4: 0.014619

Comparison: Flickr8K vs COCO

Aspect	Flickr8K	COCO
Dataset Size	~8,000 images	~118,000 images
Caption Format	Plain text (`.txt`)	Structured JSON
Max Caption Length	32 tokens	49 tokens
Vocabulary Size	7,577	27,089
Trainable Parameters	~5.0M	~6.7M
Training	8K images for 10 epochs	Initially 14K images (10 epochs), then 118K images (6 epochs)
Best BLEU-4 (Validation)	**0.0574** (Epoch 8)	**0.0486** (14K subset, Epoch 9)
Best BLEU-4 (Test)	**0.0513** (Epoch 8)	**0.0229** (Full COCO, Epoch 0)

Conclusion: Despite COCO being significantly larger, **Flickr8K yielded better BLEU-4 scores**. This is likely due to overfitting challenges and noise in COCO's diverse, large-scale data. COCO's average caption length is higher, requiring the model to capture more complex semantics, which might be harder with the same architecture.

Future Enhancements

- Try different CNNs, such as InceptionV3, ResNet50, or EfficientNet, for improved feature extraction.
- Train longer with learning rate decay or early stopping based on validation BLEU.
- Prevent overfitting using higher dropout and L2 regularization.
- Implement attention mechanisms (like in *Show, Attend and Tell*) to help the model focus on relevant image regions while generating each word.